

# Supplementary: Foundation Models and Adaptive Feature Selection: A Synergistic Approach to Video Question Answering

Sai Bhargav Rongali  
Indian Institute of Technology Bombay, India  
rongalisaibhargav002@gmail.com

Mohamad Hassan N C  
Indian Institute of Technology Bombay, India  
mohdhassannnc@gmail.com

Ankit Jha  
LNMIIT, Jaipur, India  
ankitjha16@gmail.com

Neha Bhargava  
Fractal AI Research, India  
neha.bhargava@fractal.ai

Saurabh Prasad  
University of Houston  
saurabh.prasad@ieee.org

Biplab Banerjee  
Indian Institute of Technology Bombay, India  
getbiplab@gmail.com

## A. Contents of the Supplementary Material

In this supplementary material, we present the following details:

- Section **B** provides detailed information on the dataset used in the experiments. We also discuss our implementation setup for the proposed LGQAVE.
- We study the effects of various loss functions used in the design of our proposed LGQAVE in Section **C**.
- To showcase the performance of our proposed LGQAVE qualitatively, we present visualization results in Section **D**.
- Finally, in Section **E**, we summarize the list of variables used in the proposal of LGQAVE.

## B. Dataset and Implementation details

### B.1. Dataset overview

Table 1 provides an overview of the datasets used in our experiments. These datasets span various domains and question-answering formats, each contributing to the evaluation of different aspects of video question answering (VQA). **NExT-QA** [8] focuses on causal and temporal reasoning with 5.4K videos and 48K question-answer pairs. **TGIF-QA** [2] is divided into three distinct tasks: Repetition Action, State Transition, and Frame QA, each with varying amounts of video and question data. **STAR-QA** [6], with its 5K videos and 60K questions, emphasizes situated reasoning. **Causal-VidQA** [4] pushes the boundaries of evidence-based and commonsense reasoning with

a large dataset comprising 26.9K videos and 161.4K questions. Lastly, **MSRVTT-QA** [9], focusing on visual recognition, provides an extensive dataset of 10K videos and 244K question-answer pairs. These datasets cover a broad spectrum of reasoning tasks and question-answering structures, which ensures the robustness and generalizability of our model.

### B.2. Implementation details

We processed each video by decoding it into frames and selecting 32 frames per video. These frames were split into 8 clips, each containing 4 frames, as described in [7]. To extract features, we used the pre-trained CLIP model ViT-B/32 [5], setting the embeddings to 100 tokens per frame and padding with empty tokens if needed. Object detection was performed using the MiniGPT model [10], which produced up to 10 graphs per clip, with any unused graphs left empty. The graph model has two layers with hidden states of size 512, and the transformer module uses one layer with 8 self-attention heads. For edge transformations in the Q-DGT, the self-attention heads are reduced to 5. Frames were selected for further processing based on a cross-attention score greater than 0.4 ( $\beta > 0.4$ ). We used a decay factor of  $\gamma = 0.9$  when computing the final feature representation,  $\mathcal{F}_{final}$ . In multiple-choice QA, wrong options were used as negative samples, while in open-ended QA, negative samples included answers from other questions and difficult negatives from the same category. The model was regularized with a parameter  $\tilde{\lambda} = 1$  and trained using the Adam optimizer [3]. We started with a learning rate of  $5 \times 10^{-5}$ , which decreased over time using a cosine annealing schedule. The batch size was set to 64, and the model was trained

Table 1. Overview of the datasets used in the experiments.

| Dataset            | #Videos/#QAs   | Train          | Val          | Test          |
|--------------------|----------------|----------------|--------------|---------------|
| NExT-QA [8]        | 5.4K / 48K     | 3.8K / 34K     | 0.6K / 5K    | 1K / 9K       |
| TGIF-QA [2]        | 91.8K / 134.7K | 79.2K / 112.6K | -            | 12.5K / 22.2K |
| ActivityNet-QA [1] | 5.8K/58K       | 4.64K/46.4K    | -            | 1.16K/11.6K   |
| STAR-QA [6]        | 5K / 60K       | 3K / 46K       | 1K / 7K      | 1K / 7K       |
| Causal-VidQA [4]   | 26.9K / 161.4K | 18.8K / 112.7K | 2.7K / 16.0K | 5.4K / 32.6K  |
| MSRVTT-QA [9]      | 10K / 244K     | 6.5K / 159K    | 0.5K / 12K   | 3K / 73K      |

for up to 30 epochs, depending on the dataset.

### C. Ablation Study on Loss Functions

We conducted a detailed ablation study to assess how different components of our composite loss function affect performance in video question answering (VQA) tasks. The plots in Figure 1 show the performance across four major datasets: NextQA, CausalQA, MSRVTT, and StarQA, using different configurations of the loss functions. As described in the main paper, the composite loss function consists of two key components, i.e.,

a)  $L_{vq}$ , captures the direct interaction between the video and the question.

b)  $L_{vqa}$ , accounts for the multi-modal interaction between the video, the question, and the multiple-choice options or the answer in an open-ended scenario. To balance the contributions of these components, we introduce a regularization factor  $\lambda$ . This leads to the combined loss function:

$$L = L_{vqa} + \lambda L_{vq}. \quad (1)$$

The bar graph shows the performance of the model on the following configurations:

- $L_{vq}$  **alone**: Represented in cyan, this configuration uses only the video-question interaction term.
- $L_{vqa}$  **alone**: Represented in light blue, it captures the interaction between the video, question, and multiple-choice options.
- $L_{vqa} + \lambda L_{vq}$ : Represented in blue, this configuration combines both loss terms with a balancing parameter  $\lambda$ .

From the results shown in Figure 1, we observe that our proposed LGQAVE performs the worst when trained only with  $L_{vq}$  loss across all datasets. In contrast, training with  $L_{vqa}$  loss, which captures the multimodal interaction between the video and the question, shows better performance compared to using only  $L_{vqa}$  loss. Specifically, the performance improves significantly with  $L_{vqa}$ , especially on the CasualQA and StarQA datasets. This indicates that interactions between the video, question, and multiple-choice options are crucial for accurate question answering.

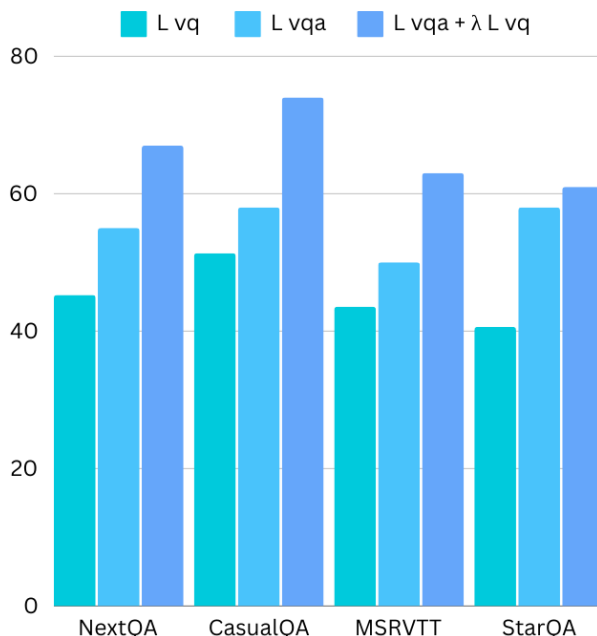


Figure 1. Effect of loss functions on our LGQAVE model.

We further combine both losses,  $L_{vq}$  and  $L_{vqa}$ , using the regularization factor  $\lambda$ , as described in Equation 1. This combined loss function achieves the highest accuracy on the NextQA and MSRVTT datasets, highlighting the complementary benefits of both loss terms. Our ablation study demonstrates the effectiveness of using a composite loss function that integrates both video-question interaction and the more complex multimodal interactions. Incorporating  $L_{vqa}$  significantly enhances model performance, while  $\lambda$  helps balance the contributions of  $L_{vq}$  and  $L_{vqa}$ , refining the overall results.

### D. Additional visual results

We showcase the qualitative results of our proposed LGQAVE and compare it with state-of-the-art methods in Figure 2. These results demonstrate the model’s ability to handle a wide range of video questions, from basic recognition tasks to more advanced reasoning challenges. For

instance, as illustrated in Figure 2, when asked about a scenario involving a man in a black jacket climbing a hill, the model successfully identified the relevant video segment depicting the action and generated the correct answer. This highlights the capability of model to recognize the video context and the specific action relevant to the question.

In another example, when presented with a question about a trainer’s actions, the model responded with ”Feeding the turtles.” This response reflects the model’s summarization of the video, likely due to the turtles’ activity being more prominent over time. However, in reality, the trainer is seated farther away, feeding a dog. This scenario illustrates the model’s growing proficiency in understanding questions, though it occasionally prioritizes dominant visual cues over subtle actions. Moreover, the model effectively handled complex, multi-object, and multi-action scenarios. For example, in a video featuring a woman speaking, a dog sitting, and another woman eating, the model accurately selected the pertinent segments to answer the question. This reinforces the model’s ability to reason across multiple events and objects, successfully detecting and interpreting simultaneous actions.

These visual results confirm the model’s improved question comprehension and its ability to provide accurate answers based on the relevant temporal segments. They also support the quantitative improvements observed across datasets, demonstrating a robust understanding of video-question interactions. We further discuss the qualitative effects of different modules incorporated in our LGQAVE model in Figure 3. This includes modules such as the sampling module, complete graph, question-aware graph, global features, and local features. We observe that the sampling module, question-aware graph, and global and local features generate more relevant answers for the video and question queries compared to other settings.

## E. Table of variables

Table 2 provides a comprehensive list of the key variables used in this paper. The ”Description” column outlines the specific roles and applications of each variable within our model, offering clarity on their function and relevance.

## References

- [1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2
- [2] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1, 2
- [3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [4] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 1, 2
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1
- [6] Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. *ArXiv*, abs/2405.09711, 2024. 1, 2
- [7] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829, June 2024. 1
- [8] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1, 2
- [9] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 1, 2
- [10] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Miniqt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

| Variable                   | Description  |
|----------------------------|--|
| $V_i^t$                    | The video frame at the $t^{\text{th}}$ time step for the $i^{\text{th}}$ instance. Dimensions are $H \times W \times 3$ , where $H$ is height and $W$ is width.  |
| $\mathcal{T}_i$            | The total number of frames in the $i^{\text{th}}$ video instance.  |
| $f_v$                      | Frozen CLIP image encoder used to extract visual features from frames.   |
| $\mathbf{E}_i^t$           | Visual features extracted from $V_i^t$ using $f_v$ . Dimension is $\mathbb{R}^{\mathcal{N} \times \mathcal{C}}$ , where $\mathcal{N}$ is the number of patches and $\mathcal{C}$ is the feature dimension. |
| $p$                        | The patch size for splitting the image into patches.   |
| $\mathcal{N}$              | Number of image patches, calculated as $\mathcal{N} = \frac{H}{p} \times \frac{W}{p}$ .  |
| $\mathcal{C}$              | Embedding dimension of visual features.  |
| $\mathbf{Q}_i$             | Text-guided query representation for the question $Q_i$ , extracted using a pre-trained RoBERTa model.   |
| $\mathcal{M}$              | Number of query tokens in the question representation.   |
| $\tilde{\mathbf{E}}_i^t$   | Projected visual features after passing $\mathbf{E}_i^t$ through the learnable projection layer $\phi_e$ .   |
| $\tilde{\mathbf{Q}}_i$     | Projected question features after passing $\mathbf{Q}_i$ through the learnable projection layer $\phi_q$ .   |
| $s_t$                      | Cross-attention score between question $\tilde{\mathbf{Q}}_i$ and visual features $\tilde{\mathbf{E}}_i^t$ of frame $V_i^t$ , used to select relevant frames.  |
| $\beta$                    | Predefined threshold value for frame selection based on $s_t$ .  |
| $\mathcal{V}_i$            | Set of selected frames from the $i^{\text{th}}$ video, based on the cross-attention score $s_t$ .  |
| $\mathcal{B}_i^{t'}$       | Bounding boxes around objects relevant to the question in frame $V_i^{t'}$ , generated by MiniGPT-4.   |
| $F_o^{t'}$                 | Region of Interest (RoI)-aligned object appearance features for each object in frame $V_i^{t'}$ .  |
| $F_s^{t'}$                 | Spatial locations of objects in frame $V_i^{t'}$ .   |
| $F_I^{t'}$                 | Frame-level feature representing the overall context of the frame $V_i^{t'}$ .   |
| $\mathcal{G}_i^{t'}$       | Frame-specific graph for frame $V_i^{t'}$ , constructed using object bounding boxes and frame context.   |
| $A^{t'}$                   | Node set for the frame-specific graph $\mathcal{G}_i^{t'}$ .   |
| $R^{t'}$                   | Edge weights in the frame-specific graph $\mathcal{G}_i^{t'}$ , calculated using self-attention on object features.  |
| $\hat{\mathbf{Q}}$         | Masked question embedding used in the Q-DGT module.  |
| $\mathcal{F}_{local}^{t'}$ | Local representation for frame $V_i^{t'}$ , derived from the Q-DGT module.   |
| $\mathcal{F}_{global}$     | Global video representation, aggregating spatial and temporal representations from all frames.   |
| $Z_{\hat{\mathbf{Q}}}$     | Textual embeddings of the question, projected into the textual information space.  |
| $\mathcal{F}_{final}$      | Final video representation, obtained by merging global and local representations using cross-attention.  |
| $\hat{A}$                  | Predicted answer for the question, based on similarity between $\mathcal{F}_{final}$ and pre-encoded answer representations.   |

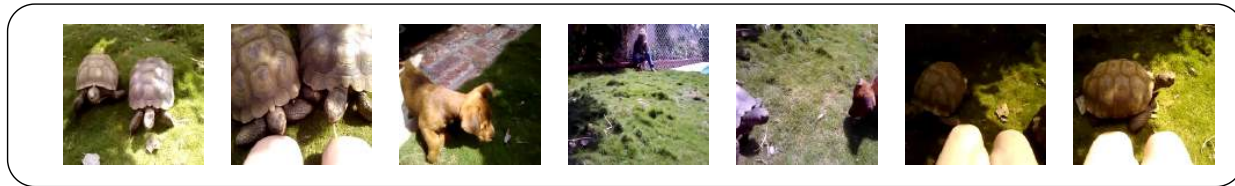
Table 2. Table of variables and descriptions used in the LGQAVE framework.



What was the man in black jacket doing ?



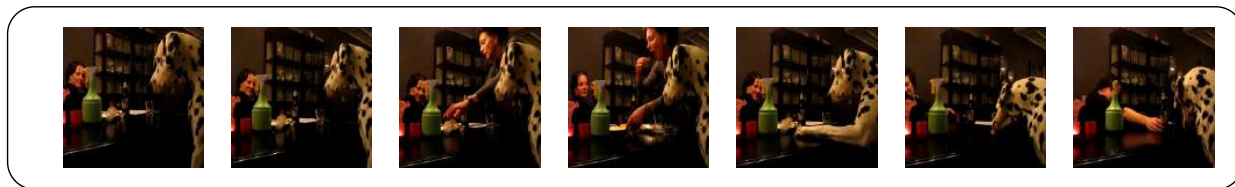
- CoVGT Climbing
- VideoChat Hiking
- VideoLlama Climbing a hill
- LGQAVE Spectating the climber



What is the trainer doing initially?



- CoVGT Sitting
- VideoChat Speaking
- VideoLlama Feeding the turtles
- LGQAVE Playing with a dog

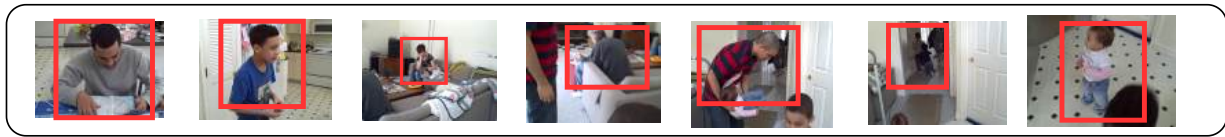



What actions are taking place in the video?



- CoVGT Speaking, walking
- VideoChat A women is speaking
- VideoLlama A women is speaking and a dog is sitting
- LGQAVE A dog is sitting , a women is speaking , another women is eating


Figure 2. Visual examples of model performance on various video question answering tasks. The model demonstrates its ability to select relevant video segments based on the question and answer accordingly, handling both simple and complex scenarios.



What is the woman doing ? 

| Sampling Module | Complete Graph | Question Aware Graph | Global Features | Local Features | Answer                             |
|-----------------|----------------|----------------------|-----------------|----------------|------------------------------------|
| <b>X</b>        | ✓              | <b>X</b>             | ✓               | <b>X</b>       | Cooking                            |
| <b>X</b>        | ✓              | <b>X</b>             | ✓               | ✓              | Watching the play                  |
| ✓               | ✓              | <b>X</b>             | ✓               | <b>X</b>       | Running after kids                 |
| ✓               | <b>X</b>       | ✓                    | <b>X</b>        | ✓              | Sitting on sofa                    |
| ✓               | <b>X</b>       | ✓                    | ✓               | ✓              | Sitting on sofa and reading a book |



What does the women in green jacket doing ? 

| Sampling Module | Complete Graph | Question Aware Graph | Global Features | Local Features | Answer  |
|-----------------|----------------|----------------------|-----------------|----------------|---|
| <b>X</b>        | ✓              | <b>X</b>             | ✓               | <b>X</b>       | Running in a playground                           |
| <b>X</b>        | ✓              | <b>X</b>             | ✓               | ✓              | Running with dog                                  |
| ✓               | ✓              | <b>X</b>             | ✓               | <b>X</b>       | Training the dog                                  |
| ✓               | <b>X</b>       | ✓                    | <b>X</b>        | ✓              | Spectating the training                           |
| ✓               | <b>X</b>       | ✓                    | ✓               | ✓              | Standing in the ground and spectating the dog run |

Figure 3. Ablation study of different components showing the strength of our model and the precise answers produced while using all the components. Inaccurate answers when missing various components show their importance.